

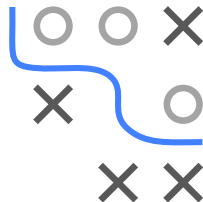
Introduction to Machine Learning

KL for ML



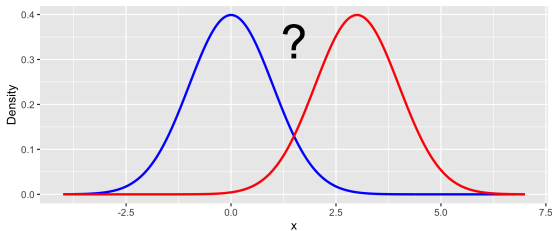
Learning goals

- Understand why measuring distribution similarity is important in ML
- Understand the advantages of forward and reverse KL



MEASURING DISTRIBUTION SIMILARITY IN ML

- Information theory provides tools (e.g., divergence measures) to quantify the similarity between probability distributions

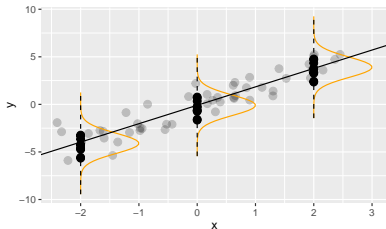


- The most prominent divergence measure is the KL divergence
- In ML, measuring (and maximizing) the similarity between probability distributions is a ubiquitous concept, which will be shown in the following.

MEASURING DISTRIBUTION SIMILARITY IN ML / 2

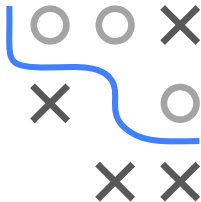
- **Probabilistic model fitting**

Assume our learner is probabilistic, i.e., we model $p(y|\mathbf{x})$ (for example, logistic regression, Gaussian process, ...).



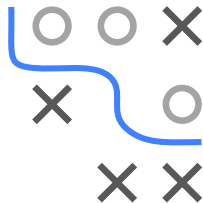
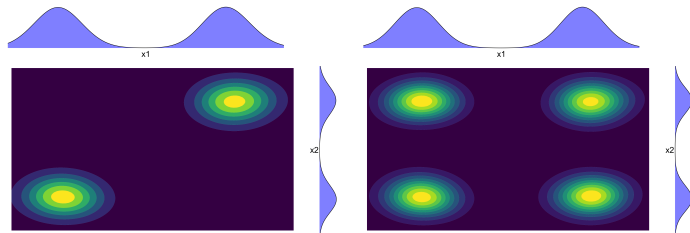
We want to minimize the difference between $p(y|\mathbf{x})$ and the conditional data generating process $\mathbb{P}_{y|\mathbf{x}}$ based on the data stemming from $\mathbb{P}_{y,\mathbf{x}}$.

Many losses can be derived this way. (e.g., cross-entropy loss)



MEASURING DISTRIBUTION SIMILARITY IN ML / 3

- **Feature selection** In feature selection, we want to choose features the target strongly depends on.



We can measure dependency by measuring the similarity between $p(\mathbf{x}, y)$ and $p(\mathbf{x}) \cdot p(y)$.

We will later see that measuring this similarity with KL leads to the concept of mutual information.

KL DIVERGENCE

Divergences can be used to measure the similarity of distributions.

For distributions p, q they are defined such that

- 1 $D(p, q) \geq 0$,
- 2 $D(p, q) = 0$ iff $p = q$.

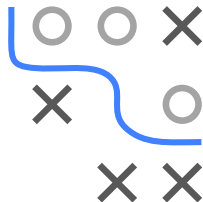
\Rightarrow divergences can be (and often are) non-symmetrical.

If the same measure dominates the distributions p, q , we can use KL.

For a target distribution p and parametrized distribution q_ϕ , we call

- $D_{KL}(p||q_\phi)$ forward KL,
- $D_{KL}(q_\phi||p)$ reverse KL.

In the following, we highlight some properties of the KL that make it attractive from an ML perspective.



KL DIVERGENCE / 2

- **Forward KL for probabilistic model fitting**

We have samples from the DGP $p(y, \mathbf{x})$ when we fit our ML model.

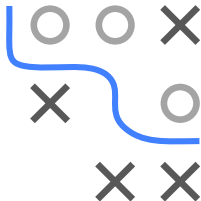
If we have a probabilistic ML model q_ϕ the expected forward KL

$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} D_{KL}(p(\cdot|\mathbf{x})||q_{\phi}(\cdot|\mathbf{x})) = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \mathbb{E}_{y \sim p_{y|\mathbf{x}}} \log \left(\frac{p(y|\mathbf{x})}{q_{\phi}(y|\mathbf{x})} \right).$$

We can directly minimize this objective since

$$\begin{aligned}\nabla_{\phi} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} D_{KL}(p(\cdot|\mathbf{x}) || q_{\phi}(\cdot|\mathbf{x})) &= \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \mathbb{E}_{y \sim p_{y|\mathbf{x}}} \nabla_{\phi} \log(p(y|\mathbf{x})) \\ &\quad - \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \mathbb{E}_{y \sim p_{y|\mathbf{x}}} \nabla_{\phi} \log(q_{\phi}(y|\mathbf{x})) \\ &= -\nabla_{\phi} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \mathbb{E}_{y \sim p_{y|\mathbf{x}}} \log(q_{\phi}(y|\mathbf{x}))\end{aligned}$$

⇒ We can estimate the gradient of the expected forward KL without bias, although we can not evaluate $p(y|\mathbf{x})$ in general.



KL DIVERGENCE / 3

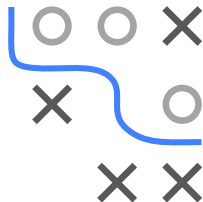
- **Reverse KL for VI**

Here, we know our target density $p(\theta|\mathbf{X}, \mathbf{y})$ only up to the normalization constant, and we do not have samples from it.

We can directly apply the reverse KL since for any $c \in \mathbb{R}_+$

$$\begin{aligned}\nabla_{\phi} D_{KL}(q_{\phi} \| p) &= \nabla_{\phi} \mathbb{E}_{\theta \sim q_{\phi}} \log \left(\frac{q_{\phi}(\theta)}{p(\theta)} \right) \\ &= \nabla_{\phi} \mathbb{E}_{\theta \sim q_{\phi}} \log \left(\frac{q_{\phi}(\theta)}{p(\theta)} \right) - \underbrace{\nabla_{\phi} \mathbb{E}_{\theta \sim q_{\phi}} \log c}_{=0} \\ &= \nabla_{\phi} \mathbb{E}_{\theta \sim q_{\phi}} \log \left(\frac{q_{\phi}(\theta)}{c \cdot p(\theta)} \right).\end{aligned}$$

\Rightarrow We can estimate the gradient of the reverse KL without bias (even if we only have an unnormalized target distribution)



KL DIVERGENCE / 4

The asymmetry of the KL has the following implications

- Forward KL $D_{KL}(p\|q_\phi) = \mathbb{E}_{\mathbf{x}\sim p} \log \left(\frac{p(\mathbf{x})}{q_\phi(\mathbf{x})} \right)$ is mass-covering since $p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q_\phi(\mathbf{x})} \right) \approx 0$ if $p(\mathbf{x}) \approx 0$ and $q_\phi(\mathbf{x}) \gg p(\mathbf{x})$.
- Reverse KL $D_{KL}(q_\phi\|p) = \mathbb{E}_{\mathbf{x}\sim q_\phi} \log \left(\frac{q_\phi(\mathbf{x})}{p(\mathbf{x})} \right)$ is mode-seeking (zero-avoiding) since $q_\phi(\mathbf{x}) \log \left(\frac{q_\phi(\mathbf{x})}{p(\mathbf{x})} \right) \gg 0$ if $p(\mathbf{x}) \approx 0$ and $q_\phi(\mathbf{x}) > 0$

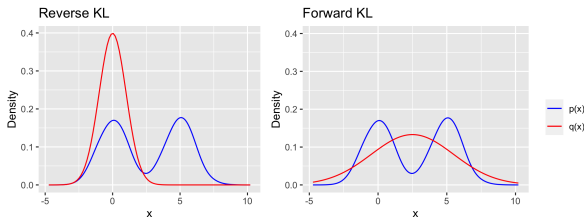


Figure: Optimal q_ϕ when q_ϕ is restricted to be Gaussian.